Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*

Tsuyoshi Kawashima[*,†], Naoki Amano[*,†,‡], Hideaki Koike[*,†], Shin-ichi Makino[†], Sadaharu Higuchi[†], Yoshie Kawashima-Ohya[†], Koji Watanabe[§], Masaaki Yamazaki[§], Keiichi Kanehori[¶], Takeshi Kawamoto[‖], Tatsuo Nunoshiba[**], Yoshihiro Yamamoto[††], Hironori Aramaki[‡‡], Kozo Makino[§§], and Masashi Suzuki[†,¶¶]

[†]National Institute of Bioscience and Human Technology, Core Research for Evolutional Science and Technology Centre of Structural Biology, 1-1 Higashi, Tsukuba 305-0046, Japan; [‡]Doctoral Program in Medical Sciences, University of Tsukuba, 1-1-1 Tennohdai, Tsukuba 305-0006, Japan; [§]Bioscience Research Laboratory, Fujiya, 228 Soya, Hadano 257-0031, Japan; [¶]DNA Analysis Department, Techno Research Laboratory, Hitachi Science Systems, 1-280 Higashi-Koigakubo, Kokubunji 185-8601, Japan; [‖]Department of Biochemistry, Hiroshima University, School of Dentistry, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan; [**]Department of Molecular and Cellular Biology, Biological Institute, Graduate School of Science, Tohoku University, Sendai 980-8578, Japan; [††]Department of Genetics, Hyogo College of Medicine, Nishinomiya 663-8501, Japan; [‡‡]Department of Molecular Biology, Daiichi College of Pharmaceutical Science, 22-1 Tamagawa-cho, Minami-ku, Fukuoka 815-8511, Japan; and [§§]Department of Molecular Microbiology, The Research Institute of Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita 565-0871, Japan

Edited by Aaron Klug, Royal Society of London, London, United Kingdom, and approved October 16, 2000 (received for review August 18, 2000)

The complete genomic sequence of the archaeon *Thermoplasma volcanium*, possessing optimum growth temperature (OGT) of 60°C, is reported. By systematically comparing this genomic sequence with the other known genomic sequences of archaea, all possessing higher OGT, a number of strong correlations have been identified between characteristics of genomic organization and the OGT. With increasing OGT, in the genomic DNA, frequency of clustering purines and pyrimidines into separate dinucleotides rises (e.g., by often forming AA and TT, whereas avoiding TA and AT). Proteins coded in a genome are divided into two distinct subpopulations possessing isoelectric points in different ranges (i.e., acidic and basic), and with increasing OGT the size of the basic subpopulation becomes larger. At the metabolic level, genes coding for enzymes mediating pathways for synthesizing some coenzymes, such as heme, start missing. These findings provide insights into the design of individual genomic components, as well as principles for coordinating changes in these designs for the adaptation to new environments.

We have determined the complete genomic sequence (composed of 1,584,799 bases) of the archaeon *Thermoplasma volcanium*. Part of this sequence is shown in Fig. 1 *A* and *C* by assembling microsquares of four colors. The genus *Thermoplasma* is unique among archaea, as it is a candidate for the origin of eukaryotic nuclei in the endosymbiosis hypothesis (1) and it is adaptable to aerobic as well as anaerobic environments (2). However, the focus of this paper is the optimum growth temperature (OGT) of *T. volcanium*, 60°C (2), which is the lowest among the archaea whose complete genomic sequences have been determined.

A number of genomic factors correlating with OGT are identified in this paper by systematically comparing this genomic sequence with the genomic sequences of seven other archaea possessing higher OGTs. In the figures, a single numbering scheme is used for referring to these archaea; 1 for *Pyrococcus furiosus* (http://www.genome.utah.edu/), 2 for *Pyrococcus* OT3 (3), 3 for *Pyrococcus abyssi* (http://www.genoscope.cns.fr/), 4 for *Aeropyrum pernix* (4), 5 for *Methanococcus jannaschii* (5), 6 for *Archaeoglobus fulgidus* (6), 7 for *Methanobacterium thermoautotrophicum* (7), and 8 for *T. volcanium*.

Materials and Methods

Determination of the New Genomic Sequence. Fragments of the genomic DNA of the strain *T. volcanium* GSS1 (Japan Collection of Microorganisms number 9571) were cloned and sequenced. The sequences of 138 DNA fragments were assembled into 30 contigs. The remaining gaps were bridged by DNA fragments constructed using the PCR. The average repetition in sequencing the same base positions was 13-fold. The overall completeness of the determined genomic sequence was confirmed by using another set of 140 fragments that altogether covered 88% of the genome. The termini of these fragments were sequenced so that their positions in the determined genomic sequence could be identified. The numbers of bases found in the genomic sequence between these 140 sets of termini were then compared with the sizes of the fragments estimated by their electrophoretic migration. The details of the sequence determination process have been reported (8). Biological analysis of the sequence is reported in this paper.

Triplet Periodicity in Gene-Coding Region. The factor that we have found to be most efficient for the identification of genes of *T. volcanium* is the preference for positioning different types of nucleotides in each of the three codon phases. This periodicity is observed uniquely in the gene-coding regions of various organisms (9, 10), although its implication is not clear (11, 12). Given a set of standard genes, the average frequencies of each of the four bases in each of the three phases can be determined. These values compose a model so that candidates that fit the model most closely are most likely to be protein genes themselves. For an A base in phase 1, the average frequency with which A occurs in the first phase in the standard genes is a measure of how well this position fits the model. Repeating this process and summing at all positions results in a score for how well the model fits, which is essentially a measure of the "gene likelihood" of the candidate.

Genes selected as the standard set were those essential for the survival of this organism: RNA polymerase, tRNA synthetases, ribosomal proteins, etc. These genes were identified manually by using their high sequence homologies to proteins of other organisms having these functions. For each gene candidate $\Sigma F^{C}_{nN} x F^{S}_{nN}$ for n = 1–3 and N = A, T, G, C was

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: OGT, optimum growth temperature.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AP000991–AP000996).

[*]T. Kawashima, N.A., and H.K. contributed equally to this work.

[¶¶]To whom reprint requests should be addressed. E-mail: sbcsuzuk@nibh.go.jp.

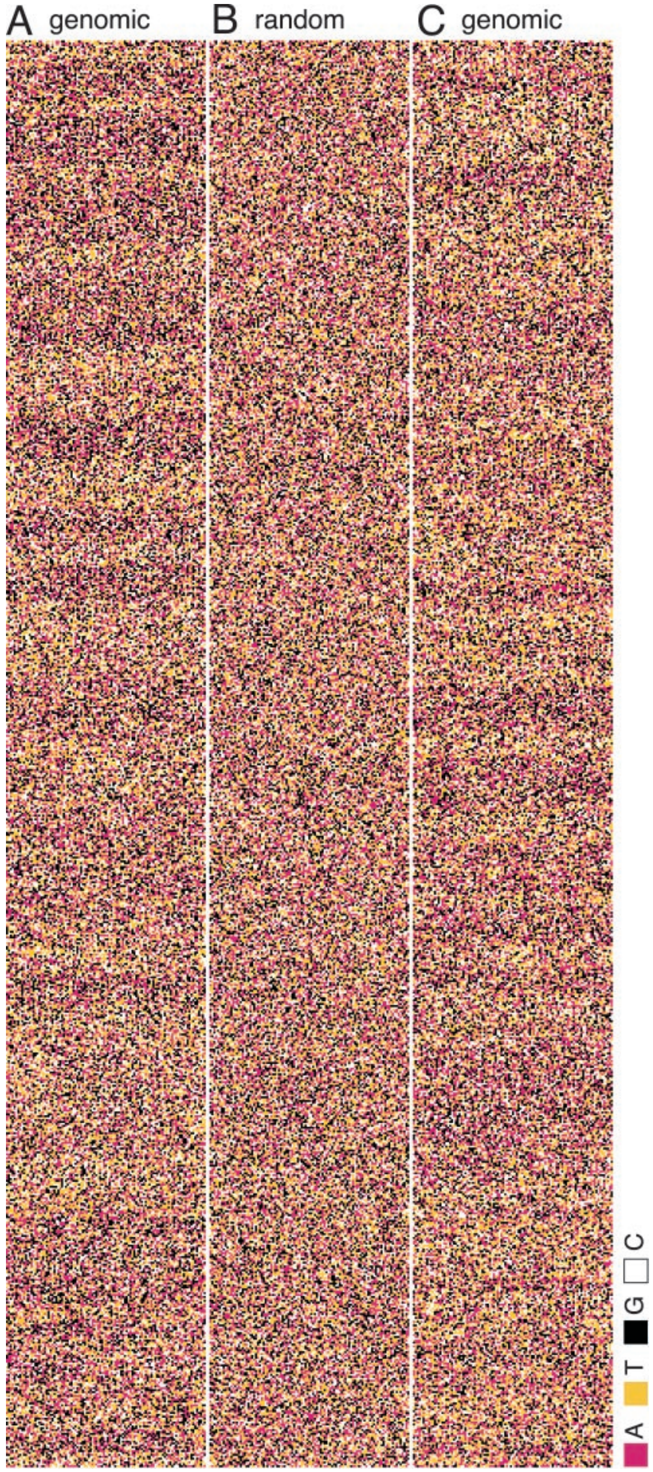


Fig. 1. Part of the genomic sequence of *T. volcanium* presented by assembling microsquares of different colors depending on the nucleotide types (*A* and *C*) in comparison with a random sequence possessing the same A/T/G/C content (*B*). Each column is composed of 750 lines from top to bottom, and each line is an assembly of 105 microsquares from left to right. Note that this figure is made so that the original sequence can be reconstituted by magnifying the figure. In *A* and *C*, gene-coding regions can be visualized as stripes in red and black, as purines are more frequent than pyrimidines in these regions (see text).

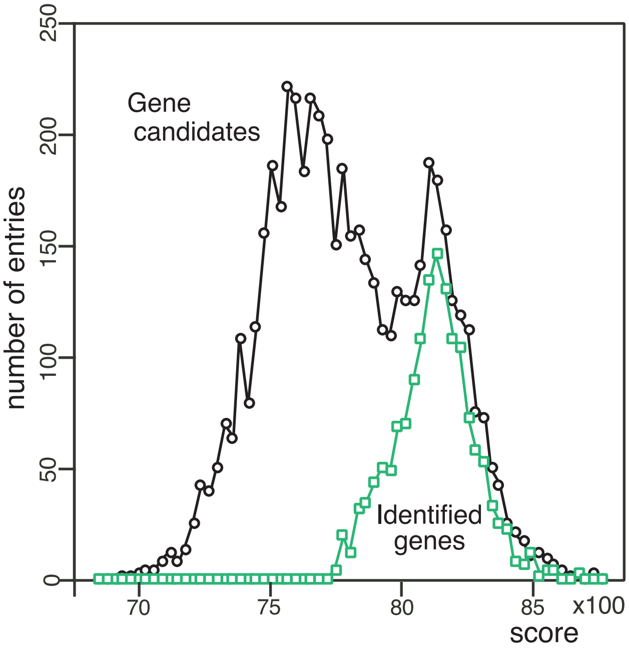


Fig. 2. Protein gene candidates of *T. volcanium* totaling 5,321 (black) and its subset of 1,524 candidates identified as genes (green) compared in the light of the triplet periodicity. Frequencies of each of the four types of bases found in each of the three codon phases were calculated for individual gene candidates. The closeness of these frequencies to those found in the standard set of genes was evaluated by the scheme described under *Materials and Methods*, and the resultant scores were plotted.

calculated, where, e.g., F^{C}_{1A} was the frequency in the candidate of A in phase 1, while, e.g., F^{S}_{1A} was a similar average frequency in the standard genes (Fig. 2). While analyzing the scores, a normal distribution was assumed for the larger subpopulation of false candidates; a threshold value was chosen so that all of the real genes, and the smallest number of false candidates, would clear the threshold.

Transcription and Translation Signals. Important measures used for further examining candidates possessing high levels of triplet periodicity were scores for the likelihood of the nucleotide sequences upstream of candidates to function as transcription or translation signals (13, 14). For each of these signal types, a scoring system similar to that used for measuring the triplet periodicity was used by considering the frequencies at individual positions as well as the frequencies of dinucleotide combinations. Prediction of transcription and translation signals is essential for the optimum determination of protein sequences, as almost all protein genes have multiple numbers of possible start codons.

The Shine–Dalgarno sequence GGAGGTGA, which is complementary to the terminus of 16S rRNA, is the consensus archaeal translation signal. Deviation from the consensus among sequences identified as signals for 167 protein genes commonly found in all of the eight thermophilic archaea decreases with increasing OGT. Consequently, the estimated average binding energy (15) between 16S rRNA and identified signals increases so that binding can be maintained at higher temperatures, with a correlation coefficient, *r*, of 0.78 along the line OGT = 16.3ΔG-77.2. For initiation of transcription, TATA-binding protein and transcription factor B are required. The transcription signals identified upstream of putative genes in *Pyrococcus* OT3 (OGT of 98°C) were closer to AAAATTTAT$\underline{\text{AAA}}$ (16), whereas those identified in *T. volcanium* in this study were closer to AAAATTTAT$\underline{\text{ATA}}$. The consensus sequence of *Pyrococcus* OT3 has a higher equilibrium binding constant to TATA-binding protein, as the consensus sequence of *T. volcanium* has a higher

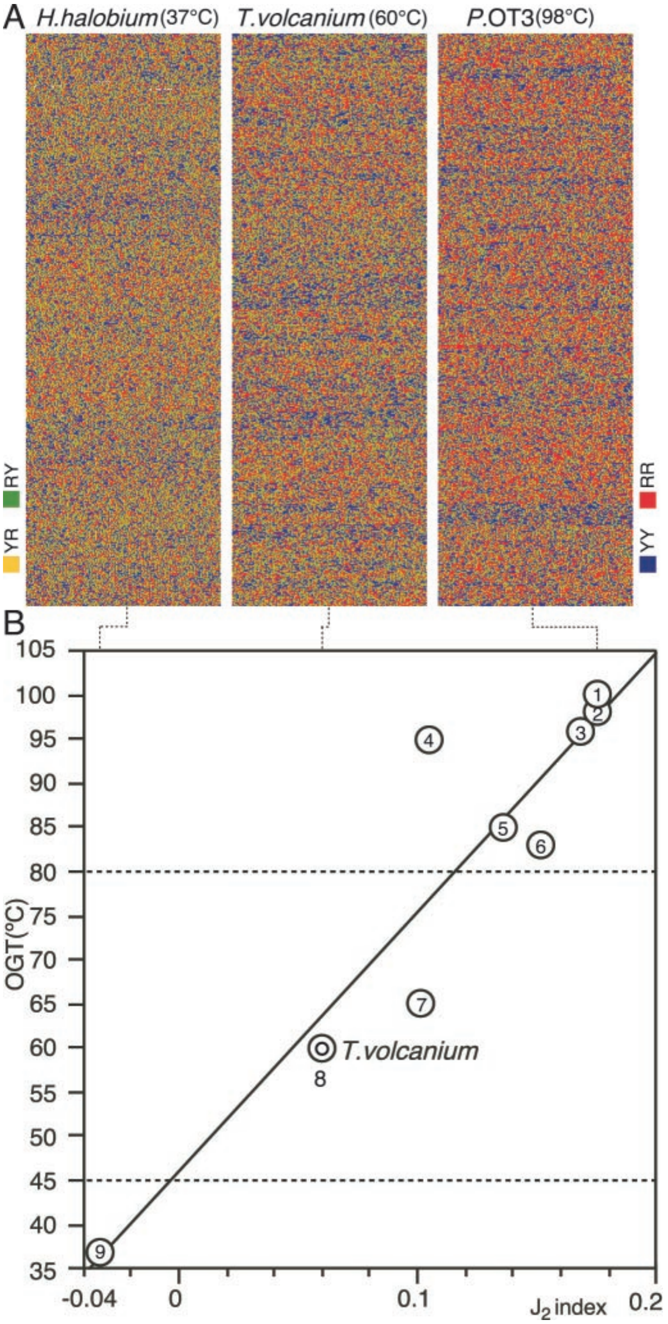


Fig. 3. Dinucleotide combination and OGT. (*A*) Part of genomic sequences represented by 95,400 (180 × 530) microsquares each. Microsquares are colored differently, depending on the purine/pyrimidine combination of dinucleotides (YY, RR, YR, RY). (*B*) The J_2 index, $\Sigma(F_{YY} + F_{RR} - F_{YR} - F_{RY})$, plotted versus OGT. The highest correlation ($r = 0.93$) is found along the line OGT = $293J_2 + 46$.

dissociation rate constant, probably because it is too flexible to maintain the complex with TATA-binding protein (T. Yamasaki, H.K., and M.S., unpublished observations).

Further details of the gene identification will be discussed elsewhere (N.A. and M.S., unpublished results). Detailed information on identified genes will be released to the ARCHAIC database (http://www.aist.go.jp/RIODB/archaic/).

Results and Discussion

Dinucleotide Combination in Genomic Sequences. The first factor whose correlation with OGT is discussed occurs at the level of the nucleotide sequence itself. In Fig. 3*A* partial genomic sequences of *T. volcanium* and *Pyrococcus* OT3 are represented as assemblies of microsquares. For comparison, a random assembly of the sequences of fragments of the mesophilic archaeon *Halobacterium halobium* is also represented (collected from GenBank http://www.ncbi.nlm.nih.gov/; number 9 in Fig. 3*B*). No complete sequence of any mesophilic archaeon has been determined.

Here, these microsquares are differentiated depending on the purine (R)/pyrimidine (Y) composition of dinucleotides: YY (TT, CC, TC, CT), YR (TA, TG, CA, CG), RY (AT, AC, GT, GC), and RR (AA, GG, AG, GA). With increasing OGT, the overall color of the presentation changes reflecting a shift from a predominance of YR and RY to one of YY and RR. This is consistent with our earlier finding (11); in an incomplete version of the genomic sequence of the hyperthermophile *Pyrococcus* OT3, the YY and RR combinations all occur with frequencies higher than those expected from a random combination of mononucleotides, whereas the YR and RY combinations all occur with frequencies lower than those expected.

To characterize this correlation mathematically, an index, J_2, has been introduced, which is defined as the subtraction of the frequency of all of the YR and RY combinations from that of all of the YY and RR combinations (Fig. 3*B*). Positive J_2 values are calculated for the sequences of all of the thermophiles, but a negative value was calculated for the mesophile *H. halobium*. A high correlation (with a correlation coefficient, *r*, of 0.93) has been found between J_2 of the nine sequences and OGT along the line OGT = $293J_2$+46.

This high correlation is consistent with an expectation that similar appropriate levels of flexibility need to be maintained by genomic DNAs for proper interaction with proteins at different OGTs. Thus, genomic DNA molecules of hyperthermophiles are expected to be least flexible. In fact, statistical study of DNA crystal structures has shown that the positioning of purine and pyrimidine rings in the YR combinations creates the smallest conformational restrictions (17–20). Indeed, DNA structures have the largest deviations at YR combinations, the DNA being largely bent at these points in many complexes with proteins. Therefore, YR combinations have been related to the flexibility of DNA. Frequencies of YR and RY become associated, because to produce a second YR, the first YR needs to be followed by an RY.

Further details of the predominance of YY/RR to YR/RY in genomes of thermophiles will be described elsewhere. In short, this predominance was found both in gene-coding regions and, similarly, in noncoding regions. In gene-coding regions, the predominance is created by the amino acid content of proteins and the codon usage for individual residues. For instance, the content of methionine decreases with increasing OGT, and this results in a decrease in the frequency of YR/RY, where the single codon of Met ATG possesses only RY and YR. Among the arginine codons, frequency of those possessing only RR rises with increasing OGT. As a consequence of all of these factors, pairs of bases positioned at the junction of codons in phases 3 and 1, and those positioned inside codons in phases 2 and 3, contribute to enhance the predominance to similar extents.

No Correlation Between Genomic G/C Content and OGT. Because an increase in the G/C content raises the stability of interactions between a pair of DNA strands (21), it is reasonable to expect high G/C contents in the genomes of thermophiles. In fact, no meaningful correlation is found between OGT and the genomic G/C content. Cations stabilize the double-stranded conformation of DNA by canceling the negative charges of phosphates (21), and increasingly higher *in vivo* K^+ concentrations have been reported for some thermophilic archaeal species with increasing OGT (22). With the K^+ concentration in hyperthermophiles of the *Pyrococcus* genus being as high as 800 mM (23), the double-stranded conformation can be maintained at temperatures close to 100°C (24).

High correlations have been found between OGT and the G/C content of tRNA and rRNA sequences of the nine archaea; $r = 0.92$ along the line OGT $= 4.9[\mathrm{G/C\%}]_{\mathrm{tRNA}} - 250$, and $r = 0.89$ along the line OGT $= 3.9[\mathrm{G/C\%}]_{\mathrm{rRNA}} - 162$. However, dependence of unfolding temperatures (TM) of RNA molecules on their G/C content is not high; TM $= 0.38[\mathrm{G/C\%}]_{\mathrm{tRNA}} + 58$ and TM $= 0.58[\mathrm{G/C\%}]_{\mathrm{rRNA}} - 27$ (calculated by using data in ref. 25). Thus, unfolding temperatures are expected to increase only by approximately 10% of the increase in OGT. Other factors such as base modification (25, 26) and the *in vivo* salt concentration need to be considered additionally.

Principles of Gene Identification. Other factors correlating with OGT have been identified by analyzing protein genes predicted to be coded in the genome of *T. volcanium*. These genes have been identified by a method essentially the same as that applied to the genome of *Pyrococcus* OT3 (refs. 16 and 27; see also *Materials and Methods*). On average, genes are coded approximately every 1,000 bases in microorganisms. Whereas approximately 1,600 genes are expected in the genome of *T. volcanium*, there exist 5,321 sections beginning with start codon combinations, ending with stop codon combinations, and being formally able to code for 50 or more amino acid residues. Therefore, precise gene identification becomes possible only by finding factors in the light of which the two subpopulations of real genes and false candidates are separated from each other as much as possible (28).

Of these candidates, 1,524 were ultimately identified as real genes. In Fig. 2, one of the factors used in our gene identification, the triplet periodicity (see *Materials and Methods*), is plotted for the total set of candidates and the subset finally identified as genes. The use of this characteristic is rationalized by the clear bimodal distribution of all candidates, separating approximately 1,600 candidates from the rest. In contrast, identified genes show a unimodal distribution, corresponding well to one of the two initial subpopulations. We take this fact as an indication of the overall precision of our gene identification.

Genes can be identified on the basis of their sequences rather than their homology to known genes, and this can be readily represented visually. In Fig. 1 *A* and *C*, gene-coding regions can be visualized as stripes in red and black, where regions are coding for genes along the strand of the sequence shown, or stripes in yellow and white, where regions are coding for genes along the complementary strand. These stripes are created because in gene-coding regions purines are more frequent than pyrimidines as a consequence of the triplet periodic combination (16, 29). The average number of bases in genes is approximately 1,000; thus, in Fig. 1 *A* and *C*, genes are expected to span multiple lines. No such stripe is seen in the same type of representation of a random combination of the four bases (Fig. 1*B*).

Bimodal Distribution of Protein pI. When high resolution two-dimensional electrophoresis was first applied to the proteins extracted from *Escherichia coli* (30), it was noticed that the majority of proteins were acidic. Later, this was complemented by systematic collection of protein sequences confirmed at the amino acid level (31). In contrast, the pI values of the proteins determined in this study to be encoded in the genome of *T. volcanium* show a bimodal distribution (Fig. 4*A*), with the two subpopulations separated by a pI gap at around 8. Proteins identified to be coded in the genome of *Pyrococcus* OT3 (27) show a similar bimodal distribution, but the basic subpopulation is larger than that of *T. volcanium* (Fig. 4*A*).

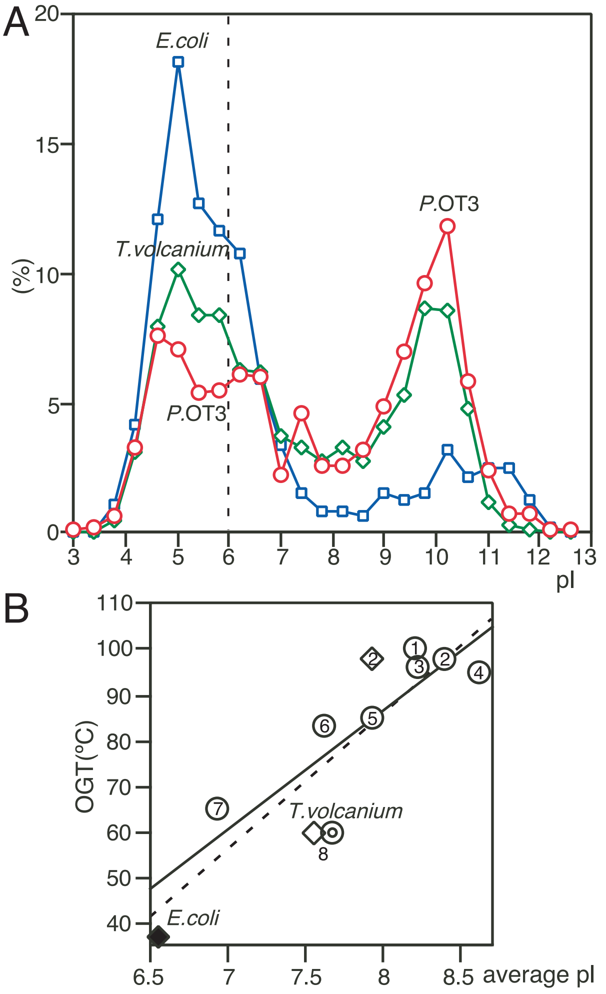


Fig. 4. Protein isoelectric point and OGT. (*A*) Distribution of isoelectric points of proteins of *Pyrococcus* OT3 (red) and *T. volcanium* (green) theoretically identified, and of *E. coli* (blue) experimentally identified (31). *In vivo* pH level of two archaea, *Thermoplasma acidophilum* (33) and *Sulfolobus acidocaldarius* (34), is indicated at 6.0. (*B*) Average isoelectric points calculated for a set of 167 proteins (○) and for the proteins in *A* (◇) plotted vs. OGT. Among the former, a high correlation ($r = 0.79$) is identified along the solid line, OGT = 26 pI − 123, whereas a higher correlation ($r = 0.88$) is identified along the broken line, OGT = 30 pI − 152, for all of the entries. Note that with increasing pI, OGTs of thermophiles change from 60–65°C (numbers 7 and 8) through 83–85°C (numbers 5 and 6) to 95–100°C (numbers 1–4) stepwise in this order, showing the correlation of the two parameters.

To confirm the relationship between pI and OGT, pI values of 167 types of proteins commonly coded in the genomes of the eight thermophiles have been compared. Within each genome, the genes of these proteins are unique, and across the different genomes high homology between the sequences of each type is found. Thus, these are reliable sets of orthologous proteins. Average pI values of sets of proteins coded in individual genomes show a high correlation ($r = 0.79$) to OGT along OGT = 26 pI − 123 (solid line in Fig. 4*B*). When the average pI values of the three protein populations in Fig. 4*A* are added to the calculation, an even higher correlation ($r = 0.88$) is found along OGT = 30 pI − 152 (broken line in Fig. 4*B*).

This increase in the average pI of 167 proteins is induced mainly by a decrease in the content of aspartic acid along the line $[\mathrm{Asp\%}] = -0.0364\mathrm{OGT} + 8.39$ ($|r| = 0.96$), and by an increase in that of lysine and arginine along the line $[\mathrm{Lys\%}] + [\mathrm{Arg\%}] = 0.0501\mathrm{OGT} + 10.53$ ($r = 0.81$).

Protein pI and OGT. The ratio of the content of amino acid residues possessing pK values closer to 7.0, His and Cys, to that of extremely acidic or basic residues in proteins, Asp, Glu, Lys, Tyr, and Arg, is very small, 1 in 8 to 1 in 10. A simple simulation can show that as a consequence of this ratio, the gap in pI at around 8 naturally occurs. When this ratio is small, the pI of a protein is kept close to neutral pH only by canceling the charges of highly acidic and highly basic residues with each other. However, incorporation of a single acidic or basic residue breaks this balance, largely shifting the protein pI to a point close to the pK of the new residue.

We have not yet formulated a concrete explanation for the correlation of the average protein pI and OGT. However, there seem to be some hints. It has been observed that proteins extracted from cells, even after dilution, become unfolded and aggregated at high temperatures (32). The smallest net charge is expected for proteins in the acidic subpopulation, as archaeal *in vivo* pH is approximately 6 (33, 34) and thus is close to the average pI of this subpopulation, approximately 5.5 (Fig. 4*A*). As unfolding of proteins is accelerated with increasing OGT, a possible strategy to minimize aggregation of proteins that are highly concentrated *in vivo* is to shift protein pI values to the basic side of the gap around 8, thereby increasing the net charge.

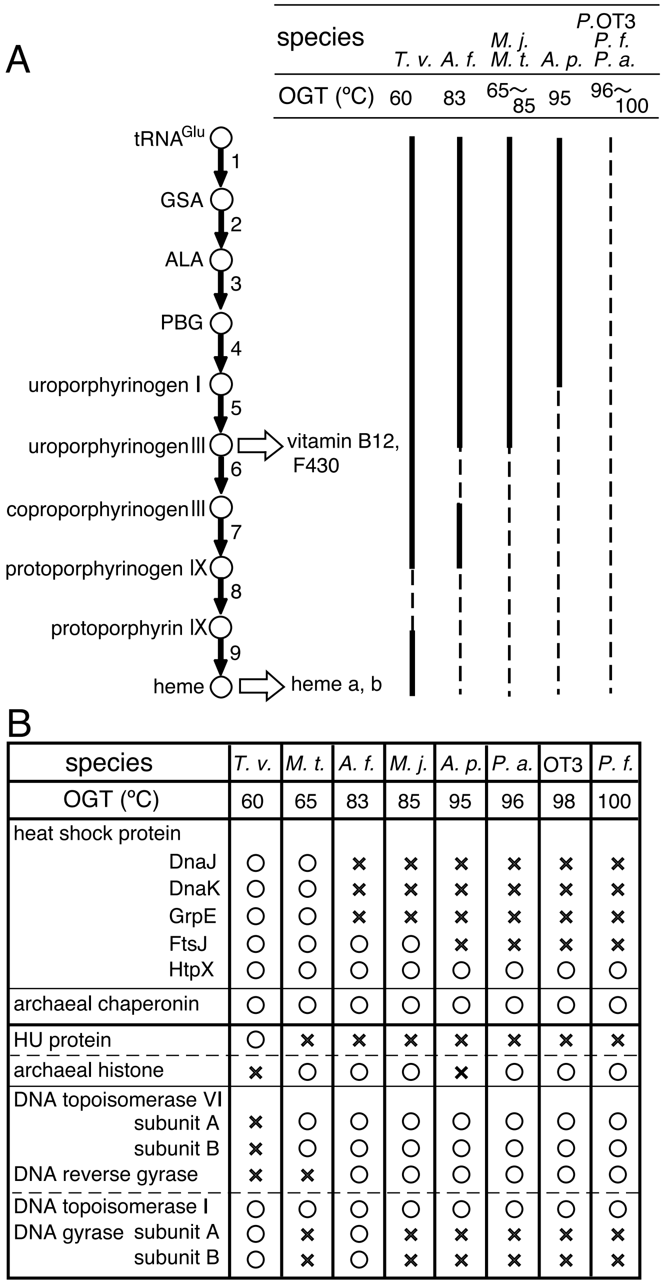
As has been reported with a eubacterium (35), lack of precision in gene identification as well as difficulty in extracting basic proteins can cause inconsistencies between proteins theoretically predicted and experimentally identified. We have found that the efficiency in extracting basic proteins is largely improved by the use of HCl. With this method, basic proteins have been extracted in increasingly larger numbers and amounts from *E. coli*, *T. volcanium*, and *Pyrococcus* OT3 in this order.

It is not simple to estimate temperature-dependent changes of protein pI. However, if the temperature-dependent changes in pK values of Asp and Glu, and of Lys and Arg, respectively, are similar to those of carboxylic acids and amines, the correlation between OGT and pI will remain.

Metabolism and OGT. Correlation with OGT is also found at the level of individual proteins. Metabolic precursors of some coenzymes, such as heme, are unstable at high temperatures (36) so that, with increasing OGT, the pathways for synthesizing these coenzymes will be significantly modified or lost. The pathway for synthesizing heme is expected to be intact in *T. volcanium*, as nearly all of the enzymes mediating this pathway have been identified (Fig. 5*A*). Methanogens surviving at higher temperatures have enzymes sufficient to constitute the first half of the pathway, up to the point where a branch for synthesizing coenzyme F430 can be entered. However, none of these enzymes is identified in *Pyrococcus* surviving at the highest OGT, suggesting that heme-related coenzymes are totally replaced by more thermostable ones. Similarly, pathways for synthesizing acetyl CoA, acyl CoA, and folic acid become incomplete with increasing OGT.

The presence of some other genes in the archaeal genomes is also related to OGT (Fig. 5*B*). Functions of some of these products are associated with superstructure formation of DNA. Hyperthermophilic archaea have enzymes for producing and relaxing positive superhelicity in DNA (37), reverse gyrase, and topoisomerase VI, respectively. These genes are missing from the genome of *T. volcanium*, which has gyrase and DNA topoisomerase I for producing and relaxing, respectively, negative superhelicity. Furthermore, the eubacterial histone-like protein HU (38) is found only in the genome of *T. volcanium*, but archaeal histone, which is closely related with eukaryotic histone proteins, is missing from this organism.

Among proteins assisting in folding of other proteins (39, 40), molecular chaperones DnaJ, DnaK, and GrpE are coded only in the genomes of two archaea of the lowest OGTs. Another heat shock protein, FtsJ, is coded only in the genomes of archaea whose OGTs are 85°C or lower. Peptidylprolyl isomerase is found coded in all of the eight archaeal genomes, whereas protein disulfide isomerase is missing from these genomes. These differences do not correspond to the phylogenic subclassification of archaea; only *A. pernix* belongs to crenarchaeota, whereas the rest belongs to euryarchaeota (see National Center for Biotechnology Information database, http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/).



species	*T. v.*	*M. t.*	*A. f.*	*M. j.*	*A. p.*	*P. a.*	OT3	*P. f.*
OGT (ºC)	60	65	83	85	95	96	98	100
heat shock protein								
DnaJ	○	○	×	×	×	×	×	×
DnaK	○	○	×	×	×	×	×	×
GrpE	○	○	×	×	×	×	×	×
FtsJ	○	○	○	○	×	×	×	×
HtpX	○	○	○	○	○	○	○	○
archaeal chaperonin	○	○	○	○	○	○	○	○
HU protein	○	×	×	×	×	×	×	×
archaeal histone	×	○	○	○	×	○	○	○
DNA topoisomerase VI								
subunit A	×	○	○	○	○	○	○	○
subunit B	×	○	○	○	○	○	○	○
DNA reverse gyrase	×	×	○	○	○	○	○	○
DNA topoisomerase I	○	○	○	○	○	○	○	○
DNA gyrase subunit A	○	×	○	×	×	×	×	×
subunit B	○	×	○	×	×	×	×	×

Fig. 5. Steps for synthesizing heme mediated by enzymes identified to be coded in different genomes (*A*, straight bold lines), and proteins involved in heat shock responses (*B*, *Upper*) and DNA superstructure formation (*B*, *Lower*). Even if an enzyme is not identified by homology, its enzymatic activity might still be produced by a protein of a yet unknown sequence. Therefore, metabolic pathways need to be discussed by carefully considering overall patterns. From this point of view, it is likely that step 8 in *A* is functional in *Thermoplasma*.

Genomic Organization and Its Components. One of the ultimate goals of genome science is to predict, on the basis of a genomic sequence alone, the overall organization from molecular to higher levels, eventually leading to possible interactions of an organism with its environment. This goal can be approached by systematic comparison of genomes of related organisms adapted to both similar and different environments, thereby identifying corresponding reflections in genomic sequences. The reflections discussed in this paper are those of adaptation to higher temperatures and can be used for the prediction of OGT, given a new archaeal genomic sequence.

Among the reflections discussed in this paper, some are more indirect than others. A possible strategy for maintaining the double-stranded conformation of genomic DNA at a high temperature is to increase the intracellular salt concentration. This would require other components to be redesigned, and this could include a shift of protein pI values to avoid bulk neutralization and, thus, aggregation. However, simply shifting pI values would not generally stabilize individual proteins. Adaptation to a new environment should often necessitate a coordinated change in the genomic organization, rather than independent modification of individual components. Higher frequency of charged residues in proteins of thermophiles has been reported (41). Yet, it is not clear whether this is a process of redesigning proteins to adapt directly to high temperatures by stabilizing protein conformations with salt bridges (41) or to do so indirectly by keeping the same conformational stability at high salt concentrations.

For adapting to the same environment, different strategies can be adopted. It is not known whether higher *in vivo* salt concentration is associated with increasing OGT of eubacteria. If the double-stranded DNA conformation needs to be stabilized at high temperature by some other factors, reflection in their nucleotide sequences would be different from what is described in this paper. Thus, for identifying reflections of adaptation to the same type of environmental changes, it is important to limit genomes compared with those that all have adopted the same type of strategies. Our focus in this paper is on archaea, and findings are not meant to be generalized without further examination of other types of organisms.

We thank Dr. L. Clowney for his critical reading of the manuscript. The work carried out at the National Institute of Bioscience and Human Technology was supported by the Core Research for Evolutional Science and Technology program of the Japan Science and Technology Corporation. The work carried out by other organizations was supported by administrative aids of Marine Biotechnology Institute and by financial aids of the New Energy and Industrial Technology Development Organization.

1. Margulis, L. (1993) *Symbiosis in Cell Evolution* (Freeman, New York), 2nd Ed.
2. Segerer, A., Langworthy, T. A. & Stetter, K. O. (1988) *Syst. Appl. Microbiol.* **10,** 161–171.
3. Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., *et al*. (1998) *DNA Res.* **5,** 55–76.
4. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., *et al*. (1999) *DNA Res.* **6,** 83–101.
5. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al*. (1996) *Science* **273,** 1058–1073.
6. Klenk, H.-P., Clayton, R. A., Tomb, J.-F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., *et al*. (1997) *Nature (London)* **390,** 364–370.
7. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., *et al*. (1997) *J. Bacteriol.* **179,** 7135–7155.
8. Kawashima, T., Yamamoto, Y., Aramaki, H., Nunoshiba, T., Kawamoto, T., Watanabe, K., Yamazaki, M., Kanehori, K., Amano, N., Ohya, Y., *et al*. (1999) *Proc. Japan Acad.* **75B,** 213–218.
9. Fickett, J. W. (1982) *Nucleic Acids Res.* **10,** 5303–5318.
10. Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. (1991) *J. Theor. Biol.* **151,** 323–331.
11. Amano, N., Ohfuku, Y. & Suzuki, M. (1997) *Biol. Chem.* **378,** 1397–1404.
12. Staden, R. (1984) *Nucleic Acids Res.* **12,** 551–567.
13. Zillig, W., Palm, P., Reiter, W.-D., Gropp, F., Pühler, G. & Klenk, H.-P. (1988) Eur J. Biochem. **173,** 473–482.
14. Brown, J. W., Daniels, C. J. & Reeve, J. N. (1989) *CRC Crit. Rev. Microbiol.* **16,** 287–338.
15. Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 9373–9377.
16. Suckow, J. M., Amano, N., Ohfuku, Y., Kakinuma, J., Koike, H. & Suzuki, M. (1998) *FEBS Lett.* **426,** 86–92.
17. Suzuki, M. & Yagi, N. (1995) *Nucleic Acids Res.* **23,** 2083–2091.
18. Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. & Dickerson, R. E. (1996) *J. Mol. Biol.* **261,** 239–254.
19. Olson, W. K. & Zhurkin, V. B. (1996) in *Biological Structure and Dynamics: Proceedings of the Ninth Conversation, State University of New York 1995*, eds. Sarma, R. H. & Sarma, M. H. (Adenine Press, New York), pp. 341–370.
20. Suzuki, M., Amano, N., Kakinuma, J. & Tateno, M. (1997) *J. Mol. Biol.* **274,** 421–435.
21. Saenger, W. (1984) *Principles of Nucleic Acid Structure* (Springer, New York).
22. Hensel, R. & König, H. (1988) *FEMS Microbiol. Lett.* **49,** 75–79.
23. De Decker, B. S., O'Brien, R., Fleming, P. J., Geiger, J. H., Jackson, S. P. & Sigler, P. B. (1996) *J. Mol. Biol.* **264,** 1072–1084.
24. Marmur, J. & Doty, P. (1962) *J. Mol. Biol.* **5,** 109–118.
25. Watanabe, K., Oshima, T., Iijima, K., Yamaizumi, Z. & Nishimura, S. (1980) *J. Biochem. (Tokyo)* **87**, 1–13.
26. Kowalak, J. A., Dalluge, J. J., McCloskey, J. A. & Stetter, K. O. (1994) *Biochemistry* **33,** 7869–7876.
27. Koike, H., Amano, N., Tateno, M., Ohfuku, Y., Suckow, J. & Suzuki, M. (1999) *Proc. Japan Acad.* **75B,** 37–42.
28. Suzuki, M. (1999) *Proc. Japan Acad.* **75B,** 81–86.
29. Makino, S., Amano, N. & Suzuki, M. (1999) *Proc. Japan Acad.* **75B,** 311–316.
30. O'Farrell, P. Z., Goodman, H. M. & O'Farrell, P. H. (1977) *Cell* **12,** 1133–1141.
31. Van Bogelen, R. A., Sankar, P., Clark, R. L., Bogan, J. A. & Neidhardt, F. C. (1992) *Electrophoresis* **13,** 1014–1054.
32. Samejima, T. & Takamiya, A. (1958) *Cytologia* **23,** 509–519.
33. Searcy, D. G. (1976) *Biochim. Biophys. Acta* **451,** 278–286.
34. Schäfer, G. (1996) *Biochim. Biophys. Acta* **1277,** 163–200.
35. Urquhart, B. L., Cordwell, S. J. & Humphery-Smith, I. (1998) *Biochem. Biophys. Res. Commun.* **253,** 70–79.
36. Daniel, R. M. & Danson, M. J. (1995) *J. Mol. Evol.* **40,** 559–563.
37. López-García, P. & Forterre, P. (1999) *Mol. Microbiol.* **33,** 766–777.
38. De Lange, R. J., Williams, L. C. & Searcy, D. G. (1981) *J. Biol. Chem.* **256,** 905–911.
39. Christensen, H., Creighton, T. E., Elöve, G. A., Fersht, A. R., Gilbert, H. F., Hartl, F. U., Hitchcock, A., Hlodan, R., Matouschek, A., Nall, B. T., *et al*. (1994) *Mechanisms of Protein Folding* (Oxford Univ. Press, Oxford).
40. Macario, A. J. L. & De Macario, E. C. (1999) *Genetics* **152,** 1277–1283.
41. Das, R. & Gerstein, M. (2000) *Funct. Integr. Genomics* **1,** 76–88.